Efficient Mining for Hadoop process with big data

Savita Suryavanshi

***Abstract*— Big data concern large-volume, complex, growing data sets that are too big. It is difficult to Big Data Mining with our current methodologies or data mining software tools, they are emerging in many important applications, such as Internet search, business informatics, and social networks, social media, genomics, and meteorology, Big Data mining grand challenge to identify the datasets and capability of extracting useful information from large datasets or streams of data The unification of multiple datasets from disparate sources in combination with advanced analytics techniques and technologies will advance problem solving capabilities, and in turn will improve the ability of predictive analysts to reveal insights that can effectively support decision making. The analysis of big data sources can be used to identify cost saving and opportunities to increase efficiency, which will directly contribute to an improvement in productivity. This can in turn help to encourage further innovations and prediction**

***Index Terms*— Big Data, data mining, heterogeneity, autonomous sources, complex and evolving associations.**

I. INTRODUCTION

Now a days big data become too big to process with our existing tool and software and main challenge is Big data Mining or Retrieving the information of different format data, various volume data and velocity data. It is difficult to data mining the big data without loss of data at micro level. Recent year data will dramatically increase due to data will collect from various sensors, applications, device, in different formats from various networks. Let consider internet data, the web page index by google were around 2 million in 1998 , it quickly reach in billion within 2 year that is at 2000 and have already exceeded 20 trillion. Information rapidly expanding or accelerating daily because of acceptance of social networking applications such as Facebook from this daily billions of data will uploaded ,Twitter from this also daily billions of data will upload ,Google Plus ,LinkedIn etc. like from many social site flood of data will Importing daily . Furthermore mobile phone becomes to get data at real-time from different ways. Vast data carries by mobile can potentially process to improve the performance of daily life. Before CDR(call data record)-based processing for billing purposes only, it can see internet of the things applications will raise the scale of data to unprecedented level .peoples are loosely connected everywhere so that millions of connected application generate large volume of data, and valuable information must discovered by improved data mining process that help improve the quality of life at short time and get valuable of information. During valuable information discovering process from the huge data, facing many challenges such that 1.Hardware and software System capability 2.Designing Algorithm business methods 3.bussines models. Adapt to the multisource, massive, dynamic Big **Data, researchers** have expanded existing data mining methods in many ways, including the efficiency improvement of **single source** knowledge discovery methods, designing a data mining mechanism from a multisource perspective, as well as the study of dynamic data mining methods and the analysis of stream data . The main motivation for discovering knowledge from massive data is improving the efficiency of single-source mining methods. On the basis of gradual improvement of computer hardware functions, resea**rchers continue to explore ways** to improve the efficiency of knowledge discovery algorithms to make them better for massive data. Because massive data are typically collected from different data sources, the knowledge discovery of the massive data must be performed using a multi source mining **mechanism. As** real-world data often come as a data stream or **a characteristic** flow, a well-established mechanism is needed to discover knowledge and master the evolution of knowledge in the dynamic data source. Therefore, the massive ,heterogeneous and real-time characteristics of multi source data provide essential differences between single source knowledge discovery and multisource data mining. proposed and established the theory of local pattern analysis, which has laid a foundation for global knowledge discovery in multisource data mining. This theory provides a solution not only for the problem of full search, but also for finding global models that traditional mining methods cannot find. Local pattern analysis of data processing can avoid putting different data sources together to carry out centralized computing. Data streams are widely used in financial analysis, online trading, medical testing, and so on. In this project system to build a stream based Big Data analytic frame work for fast response and real-time decision making. The key challenges and research issues include: - designing Big Data sampling mechanisms to reduce Big Data volumes to a manageable size for processing; - building prediction models from Big Data streams. Such models can adaptively adjust to the dynamic changing of the data. A knowledge indexing framework to ensure real-time data monitoring and classification for Big Data applications.

II. LITERATURE SURVEY

R. Ahmed and G. Karypis, Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks,[1] have recently being recognized as a powerful abstraction to model and represent the temporal changes and dynamic aspects of the data underlying many complex systems. This can help identify the transitions from one conserved state to the next and may provide evidence to the existence of external factors that are responsible for changing the stable relational patterns in these networks. This paper presents a new data mining method that analyzes the time-persistent relations or states between the entities of the

Savita Suryavanshi, Department of Computer Engineering DPCOE, Wagholi, Pune, India.

dynamic networks and captures all maximal nonredundant evolution paths of the stable relational states. [2]M.H. Alma, J.W. Ha, and S.K. Lee, Novel Approaches to Crawling Important Pages Early explain Web pages in their local repositories. In this paper, we study the problem of crawl scheduling that biases crawl ordering toward important pages. We propose a set of crawling algorithms for effective and efficient crawl ordering by prioritizing important pages with the well-known Page Rank as the importance metric. In order to score URLs, the proposed algorithms utilize various features, including partial link structure, inter-host links, page titles, and topic relevance. We conduct a large-scale experiment using publicly available data sets to examine the effecent of each feature on crawl ordering and evaluate the performance of many algorithms. [5]S. Banerjee and N. Agarwal, Analyzing Collective Behavior from Blogs Using Swarm Intelligence,

With the rapid growth of the availability and popularity of interpersonal and behavior rich resources such as blogs and other social media avenues, emerging opportunities and challenges arise as people now can, and do, actively use computational intelligence to seek out and understand the opinions of others. The study of collective behavior of individuals has implications to business intelligence, predictive analytics, customer relationship management, and examining online collective action as manifested by various flash mobs and other such events.

III. IMPLEMENTATION DETAILS

A. Problem Statement

The unification of multiple datasets from disparate sources in combination with advanced analytics techniques and technologies will advance problem solving capabilities, and in turn will improve the ability of predictive analytics to reveal insights that can effectively support decision-making. The analysis of big data sources can be used to identify cost savings and opportunities to increase efficiency, which will directly contribute to an improvement in productivity. This can in turn help to encourage further innovation.

B. Existing System

In case of existing system Static knowledge discovery methods cannot adapt to the characteristics of dynamic data streams, such as continuity, variability, rapidity, and infinity, and can easily lead to the loss of useful information. Therefore, effective theoretical and technical frameworks are needed to support data stream mining at real time as a result the unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such Big Data.

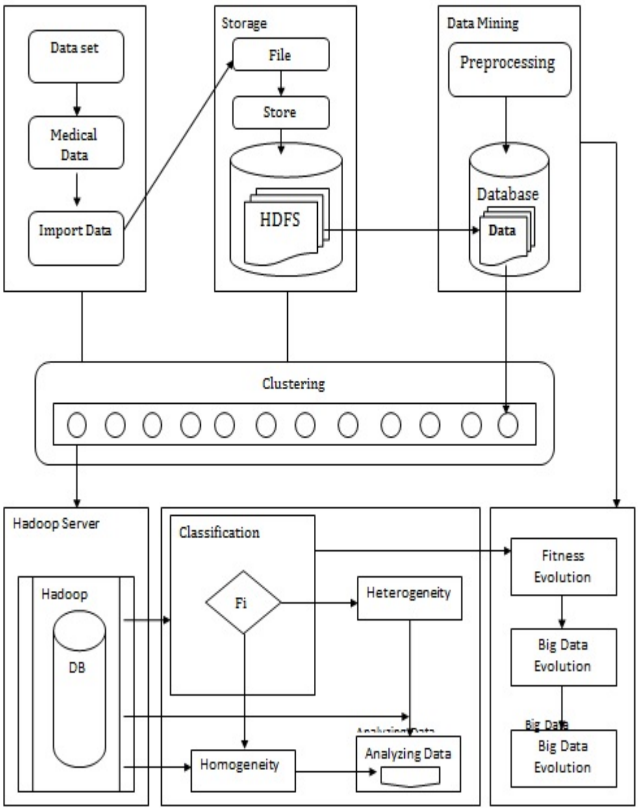
1) This paragraph is a repeat of 3.1

Please use a 9-point Times Roman font, or other Roman font with serifs, as close as possible in appearance to Times Roman in which these guidelines have been set. The goal is to have a 9-point text, as you see here. Please use sans-serif or non-proportional fonts only for special purposes, such as distinguishing source code text. If Times Roman is not available, try the font named Computer Modern Roman. On a Macintosh, use the font named Times. Right margins should be justified, not ragged.

2) Proposed System

In proposed system to build a stream-based Big Data analytic framework for fast response and real-time decision making. The key challenges and research issues include: - designing Big Data sampling mechanisms to reduce Big Data volumes to a manageable size for processing; - building prediction models from Big Data streams. Such models can adaptively adjust to the dynamic changing of the data. A knowledge indexing framework to ensure real-time data monitoring and classification for Big Data applications..

IV. PROPOSED ARCHITECTURE



A. Mathematical Model

K-means++ clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes. Given a set of observations (x1, x2, , xn), where each observation is a dimensional real vector, k-means clustering aims to partition the n observations into k (n) sets S = S1, S2, , Sk so as to minimize the within-cluster sum of squares (WCSS). In other words, its objective is to find:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}\in S_i} \|\mathbf{x}-\boldsymbol{\mu}_i\|^2$$

Where i is the mean of points in Si. Assignment step: Assign each observation to the cluster whose mean yields the least within-cluster sum of squares (WCSS). Since the sum of squares is the squared Euclidean distance, this is intuitively the To adapt to the multisource, massive, dynamic Big Data,researchers have expanded existing data mining methods in many ways, including the efficiency improvement ofsinglesource knowledge discovery methods, designing a data mining mechanism from a multi source perspective, as well as the study of dynamic data mining methods and the analysis of stream data . The main motivation for discovering knowledge from massivedata is improving the efficiency of single-source mining methods. On the basis of gradual

V. SCOPE OF WORK

To adapt to the multisource, massive, dynamic Big Data, researchers have expanded existing data mining methods in Where each is assigned to exactly one , even if it could be assigned to two or more of them. Update step: Calculate the new means to be the centroids of the observations in the new clusters**,** Since the arithmetic mean is a least-squares estimator, this also minimizes the within-cluster sum of squares (WCSS) objective. The algorithm has converged when the assignments no longer change. Since both steps optimize the WCSS objective, and there only exists a finite number of such partitioning, the algorithm must converge to a (local) optimum. There is no guarantee that the global optimum is found using this algorithm. The algorithm is often presented as assigning objects to the nearest cluster by distance. The standard algorithm aims at minimizing the WCSS objective, and thus assigns by ”least sum of squares”, which is exactly equivalent to assigning by the smallest Euclidean distance many ways, including the efficiency improvement of single source knowledge discovery methods, designing a data mining mechanism from a multi source perspective, as well as the study of dynamic data mining methods and the analysis of stream data . The main motivation for discovering knowledge from massive data is improving the efficiency of single-source mining methods. On the basis of gradual improvement of computer hardware functions, researchers continue to explore ways to improve the efficiency of knowledge discovery algorithms to make them better for massive data. Because massive data are typically collected from different data sources, the knowledge discovery of the massive data must be performed using a multisource mining mechanism. As real-world data often come as a data stream or a characteristic flow, a well-established mechanism is need efficent discover knowledge and master the evolution of knowledge in the dynamic data source in Big Data mining for high-perform.

VI. CONCLUSION

Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. At the data level, the autonomous information sources and the variety of the data collection environments, often result in data with complicated conditions, such as missing/uncertain values. Existing methods can only work in an offline fashion and are incapable of handling this Big Data scenario in real time. As a result, the unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such Big Data.

REFERENCES

[1] R. Ahmed and G. Karypis, “Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks,”Knowledge and Information Systems,vol. 33, no. 3, pp. 603-630, Dec. 2012.

[2] M.H. Alam, J.W. Ha, and S.K. Lee, “Novel Approaches to Crawling Important Pages Early,”Knowledge and Information Systems,vol. 33, no. 3, pp 707-734, Dec. 2012.

[3] S. Aral and D. Walker, “Identifying Influential and Susceptible Members of Social Networks,”Science,vol. 337, pp. 337-341, 2012.

[4] A. Machanavajjhala and J.P. Reiter, “Big Privacy: Protecting Confidentiality in Big Data,”ACM Crossroads,vol. 19, no. 1, pp. 20-23, 2012.

[5] S. Banerjee and N. Agarwal, “Analyzing Collective Behavior from Blogs Using Swarm Intelligence,” Knowledge and Information Systems,vol. 33, no. 3, pp. 523-547, Dec. 2012.

[6] E. Birney, “The Making of ENCODE: Lessons for Big-Data Projects,”Nature,vol. 489, pp. 49-51, 2012.

[7] J. Bollen, H. Mao, and X. Zeng, “Twitter Mood Predicts the Stock Market,”J. Computational Science,vol. 2, no. 1, pp. 1-8, 2011.

[8] S. Borgatti, A. Mehra, D. Brass, and G. Labianca, “Network Analysis in the Social Sciences,”Science,vol. 323, pp. 892-895, 2009.

[9] J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch.McKinSey Quarterly, 2010